Atal Bihari Vajpayee-Indian Institute of Information Technology & Management,
Gwalior - 474015

BIG DATA ANALYSIS PROJECT REPORT

" Statistical Data Analysis , Visualization & Mathematical Forecasting Of Spread of SARS-CoV-2 in India "

Submitted On : 29 - Nov - 2020

Submitted to:

Dr. Anuradha Singh

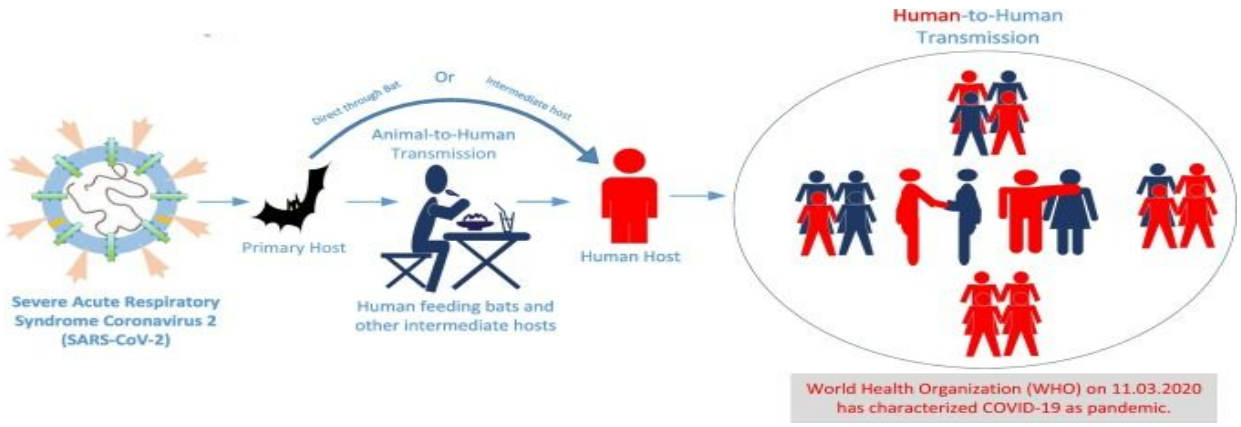
Submitted by :

Ketan Gupta (2017IMT-101)
Harshit Patel (2017IMT-039)
Ashish Kumar (2017IMT-021)
Saket Saumya (2017IMT-075)
Anshita Sharma (2017IMT-014)
Aryanshu Verma (2017IMT-019)
Kaustubh Pathak (2017IMT-100)
Dhananjai Kumar (2017IMT-033)
Suryadeepti Singh (2017BCS-031)
Aman Kumar Banka (2017IMT-094)

TABLE OF CONTENTS

1. Origin of the Proposal

Covid-19 is a disease caused by the **SARS-CoV-2** (severe acute respiratory syndrome coronavirus 2) virus making it one of the most fatal diseases that targets the human respiratory system, firstly identified in December 2019 in Wuhan, China. It spreads from one person to another through the respiratory system being exposed to the virus incoming from an infected person. The basic reproduction number, R_0 for COVID-19 has been identified to be 2-2.5, which means each person gets 2 to 2.5 people sick and if we run 10 rounds for this model, we will get around 2047 people sick with a single person. It mainly gets transmitted through droplets produced when an infected person coughs, exhales, or sneezes. Other people can also get the infection by breathing the virus inside if they are nearby of someone who has COVID-19, or by touching a contaminated surface



and then once again touching their eyes, nose or mouth.

As of 29 November 2020, more than 62 million cases have been reported across 220 countries and territories with more than 1,459,650 deaths; more than 43 million people have recovered. Currently there are around 18 million infected patients of COVID-19 around the world. In India, COVID-19 was first detected in a student from Kerala on January 30, 2020 who returned from Wuhan, China. As of now the total infected people reported in India is around 9.3 million with more than 136,733 deaths; more than 8.8 million people have recovered. But still currently there are 454,039 active cases. India ranks 2 after the USA in most number of COVID-19 cases.

The trend followed by the spread of the virus is very puzzling and interesting to analyze and deduce the effect of various factors affecting the spread of the virus. Most of the research papers and articles focus on the COVID-19 infection in the entire India. But considering the size and diversity of our country , it would be a good idea to look at the spread of the virus in each state separately along with the entire nation. We use a modified version of the mathematical model **SEIR (Susceptible, Exposed, Infectious & Recovered)** and FaceBook Build **Prophet** model which accept Time series Data is used for predicting the infection. It was reported that the infected cases' growth rate would be controlled with the help of Multiple National Lockdown, but some uncontrolled mass level events had negatively impacted the number of infected cases. In this project we have also used exponential and polynomial regression modelling to make predictions of cases till 28 November 2020.

2. Review (Status of R&D On SARS - COV 19):

In view of the emergency of the COVID-19 outbreak, the international community is mobilizing to find effective ways to rapidly enhance the development process of ways to fight covid including vaccines and therapeutics. The current COVID-19 pandemic is unparalleled, but the global response infer on the lessons that had been learned in the past decades from other disease outbreaks.

2.13 International Status:

WHO is taking numerous steps to curb the COVID-19 spread by accelerating diagnostics, vaccines and therapeutic methods. One of its successful projects is the R&D Blueprint, which attempts to strengthen collaboration between scientists and global health practitioners. On 30 January 2020, the Director-General of the WHO announced that the outbreak of COVID-19 had reached the extent of a Public Health Emergency of International Concern(PHEIC). As a part of this announcement, world scientists have gathered to determine the present state of information on the emerging virus, to agree on important scientific issues, and to identify ways to promote and finance priority research to control this epidemic and to prepare for future plans.The meeting ended with two key objectives: the first was to improve research plans to provide services for those affected, and the second was to promote research goals in hopes of learning from the latest pandemic response and to help prepare for the next outbreak.

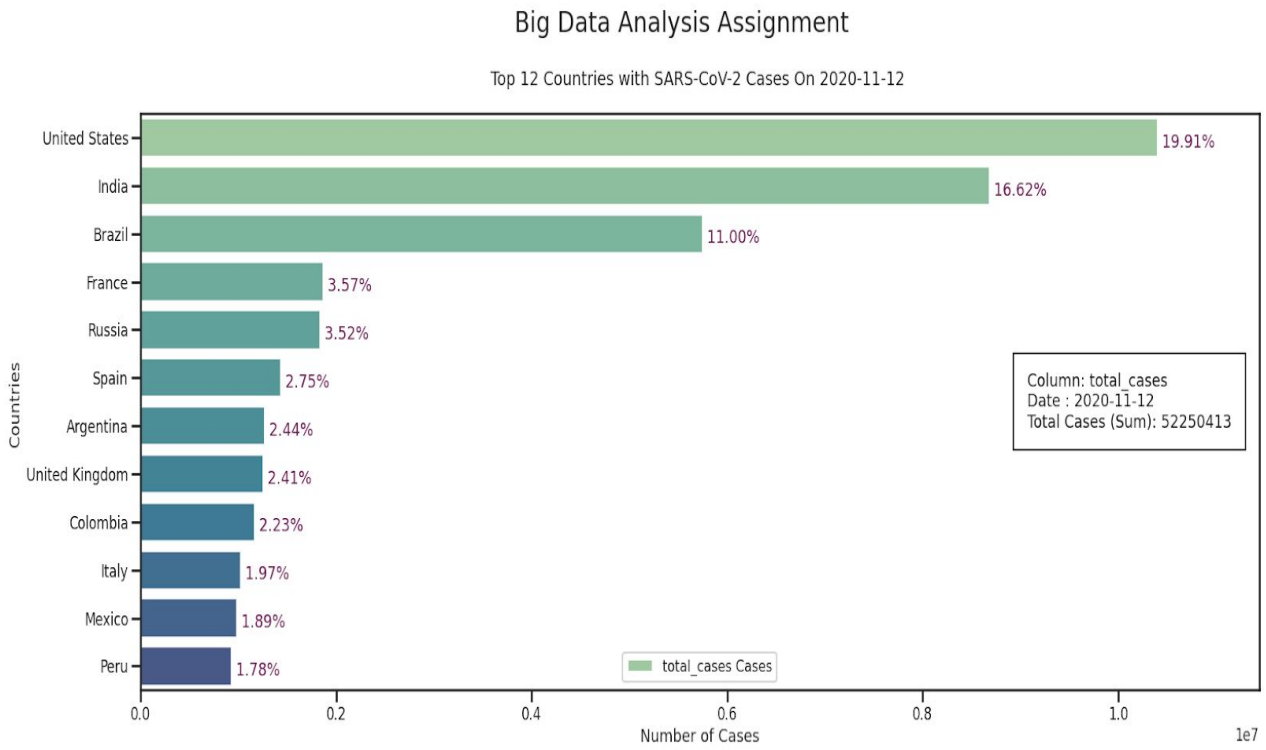
Another WHO initiative was 'Solidarity,' a multinational clinical trial of COVID-19 initiated by the WHO and its allies. The Solidarity Trial attempts to determine whether any drug increases the survival rate or decreases the need for ventilation or hospital stay. The Solidarity Trial showed that all 4 therapies based on Remdesivir, Hydroxychloroquine, Lopinavir/Ritonavir and Interferon had little to no effect on total mortality, ventilation and hospital stay-time. The Solidarity Trial is also considering other therapies for successful treatment of COVID-19. Up to now, only corticosteroids have been found to be selective against serious and dangerous COVID-19. WHO has discontinued the trial of hydroxychloroquine and lopinavir/ritonavir based on the recommendation of the International Steering Committee of the Solidarity Trial.The Committee issued a recommendation on the basis of proof of the preliminary findings of the Solidarity Trial and a summary of all the facts collected at the WHO Summit on COVID-19 Research and Innovation on 1-2 July. These intermediate findings revealed that hydroxychloroquine and lopinavir/ritonavir yield little to no reduction in the mortality rate of hospitalised COVID-19 patients relative to the standard of treatment. Therefore the Solidarity Trial investigators have now halted the trials with immediate effect. For each medication, the intermediate findings did not provide adequate proof of increased mortality. However this opinion applies solely to the execution of the Solidarity Trial of patients who are hospitalised and does not impact the future determination of other trials in non-hospitalized patients or as a pre-or post-exposure avoidance of COVID-19.As of 2 October 2020, over 12,000 patients had been enrolled in participating hospitals worldwide. The Solidarity Trial is underway in 30 countries of the 43 countries that have permission to begin trial. Overall, 116 countries have joined or expressed an interest in joining the trial. Each participating country is

a sponsor to the trial in its country and supports this endeavour, including financially. WHO is actively assisting counties with:

- identifying hospitals with willingness to participate in trial;
- training of hospital clinicians;
- shipping the trial drugs according to the needs of each participating country.

More the number of participating patients, the faster results will be produced. WHO is providing access to thousands of treatment courses to enable successful completion of trials through donations.

The current status with top 12 countries having the highest number of cases is shown in figure below:

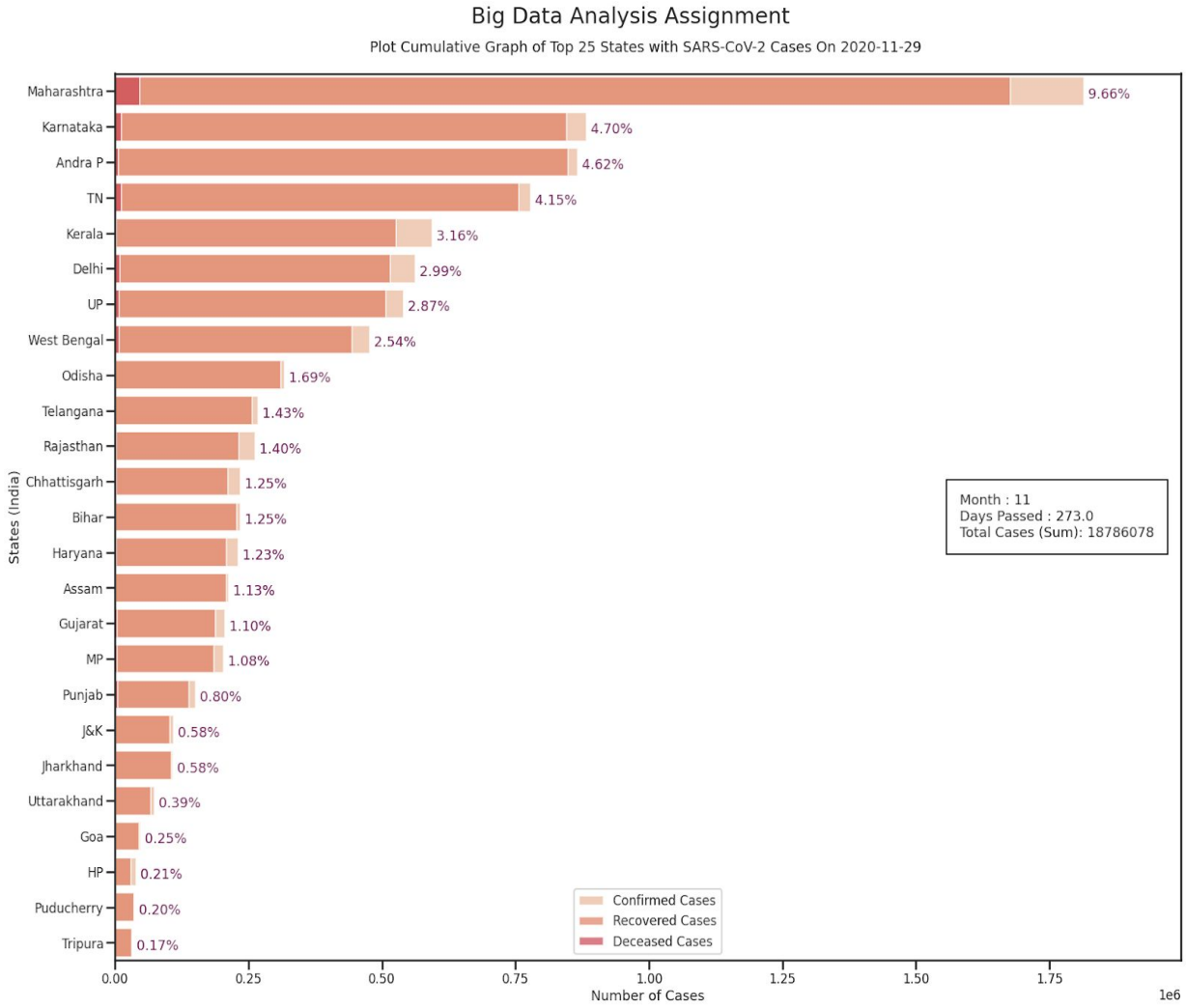


2.2 National Status:

The first COVID-19 case in India was revealed in Kerala State on January 30th, 2020. Successively, the number of cases kept on rising drastically. According to the reports of the Indian Council of Medical Research(ICMR) on Nov 28th, 2020, a total of 13,95,03,803 suspected samples had been tested in a related testing laboratory. Among them, 93,92,919 cases tested positive for SARS-CoV-2.

There have been various research and studies that have focused on the analysis and forecasting of COVID-19 situation in India. These studies have shown both long term trends as well as short term trends. These studies have used various mathematical models such as SEIR, SRID, SIR and other epidemic models. There has been forecasts using various exponential and polynomial based regression techniques and advanced forecasting methods such as ARIMA. Various machine learning and deep learning models have been used to predict and analyse the spread of COVID-19 infection in India. There have been research to study the work done by doctors and health workers.

Current status of reported positive COVID-19 cases in India(State-wise).



S. No	State	Total Cases	Active	Discharge	Death
1.	Maharashtra	18,14,515	90,965	16,76,564	46,986
2.	Karnataka	8,82,608	24,776	8,46,082	11,750
3.	Andhra Pradesh	8,67,063	11,571	8,48,511	6,981
4.	Tamil Nadu	7,79,046	11,073	7,56,279	11,694
5.	Kerala	5,93,957	64,964	5,26,797	2,196
6.	Delhi	5,61,742	36,578	5,16,166	8,998
7.	Uttar Pradesh	5,39,899	25,243	5,06,938	7,718
8.	West Bengal	4,77,446	24,537	4,44,587	8,322
9.	Odisha	3,17,789	5,510	3,10,549	1,730
10.	Telangana	2,69,223	10,490	2,57,278	1,455
11.	Rajasthan	2,62,805	28,751	2,31,780	2,274
12.	Chhattisgarh	2,34,725	20,978	2,10,917	2,830
13.	Bihar	2,33,572	5,380	2,26,939	1,253
14.	Haryana	2,30,713	19,916	2,08,422	2,375
15.	Assam	2,12,483	3,313	2,08,422	2,375
16.	Gujarat	2,06,714	14,762	1,87,969	3,953
17.	Madhya Pradesh	2,03,231	14,981	1,85,013	3,237
18.	Punjab	1,50,805	7,834	1,38,206	4,765
19.	Jammu and Kashmir	1,09,383	5,112	1,02,591	1,680
20.	Jharkhand	1,08,786	2,154	1,05,669	963
21.	Uttarakhand	73,951	4,876	67,861	1,214
22.	Goa	47,689	1,348	45,655	686
23.	Himachal Pradesh	38,977	8,574	29,780	623
24.	Puducherry	36,902	519	35,774	609
25.	Tripura	32,674	649	31,655	370

Potential treatment initiatives and approaches need to be developed against COVID-19 virus as it is proving to be a major cause of death, and it is also having unfortunate socio-economic effects, that are continually aggravated. India is tackling the said problem in an efficient way. Firstly, India is taking all the necessary preventive measures to reduce the viral transmission. Secondly, ICMR together with Ministry of AYUSH has developed guidelines to be used as conventional preventive and treatment strategies to increase immunity against COVID-19. The recent report from the director of ICMR stated that India would undergo randomized controlled trials using convalescent plasma of completely recovered COVID-19 patients. India has already much experience in special medical/pharmaceutical industries with production facilities, and as a result of that the Gov. of India has set fast-tracking research to develop rapid diagnostic test kits also known as the RT-PCR test kits and vaccines that too at low cost. The technological aspect, the Union Health Ministry has launched a mobile application called "**Aarogya Setu**" that works on all major platforms like Android and IOS. This application forms a user database for establishing a network which can provide awareness that can notify people and the concerned authorities about possible COVID-19 victims and their traces.

2.3 Importance of the proposed project in the context of current status

Many of the models used for segmentation or forecasting started to fail when traffic and shopping patterns changed, supply chains were interrupted, and borders were locked down. But since, now the lockdown in most of the countries have been removed at most of the places, the models proposed from now onwards are expected to perform better in most of the cases.

Since, we all know the impact of the COVID-19 on the economy and financial condition of countries. Example is the decrease in the GDP of developed and developing countries due to lockdown followed. In such a case, the better allocation and management of the resources of the countries becomes a very important matter. Since, the countries can become better prepared for the upcoming expected number of the infected number of cases. So, As COVID-19 surges across the country, data and predictive analytics are being used for better allocating resources with regards to testing, personal protective equipment, medications and more.

As we studied earlier how the researchers and scientists from different countries came under a single roof to solve the ongoing problems that the pandemic has proposed. Similarly, the machine learning experts have come along with the Data Analysts and tried to provide the data of the pandemic as accurately as possible. So, the Data Analysis of such an important matter of concern will also help other researchers around the globe to collaborate and take advantage of the results and conclusions.

Following the same path, we have tried modified versions of mathematical models SEIR (Susceptible, Exposed, Infectious & Recovered) and FaceBook Build Prophet for the prediction of the number of Infected and Recovered cases and employed them on. SEIR Model is a Special

Pandemic Model Most suited For Disease Like Corona itself. Due to its high Spread Rate and Ease to Infect Susceptibles make it most ideal for this situation. As for the Difference from CoronaVirus we modified the Equations that take care of the fact that Cured Patients can be reeffected by Corona Virus.

The Data sets available for the ongoing academic. We also tried to draw the comparative results between the Actual Number of Cases and the Predicted Number of Cases. All the results have been provided in the form of tools which are the most easily interpretable which are the Graphs. These graphs can easily provide other researchers the view of the pandemic and that too in the summarised formats.

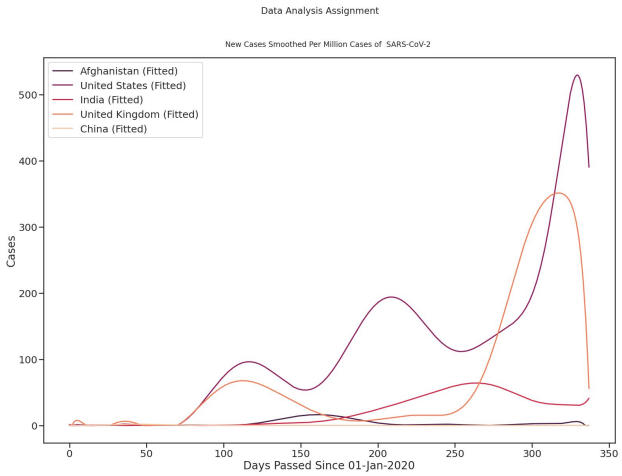


Fig 1. Confirmed Cases In Per Million

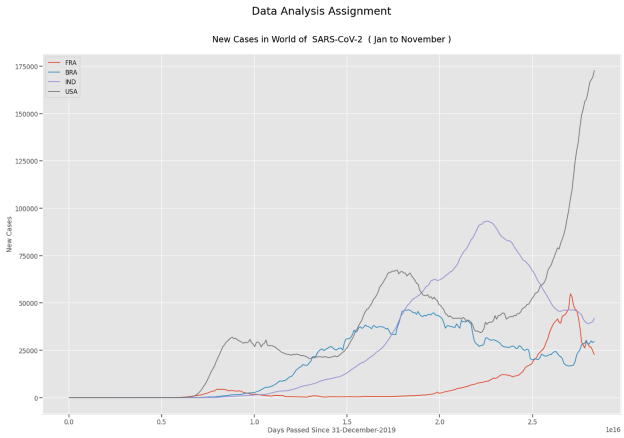
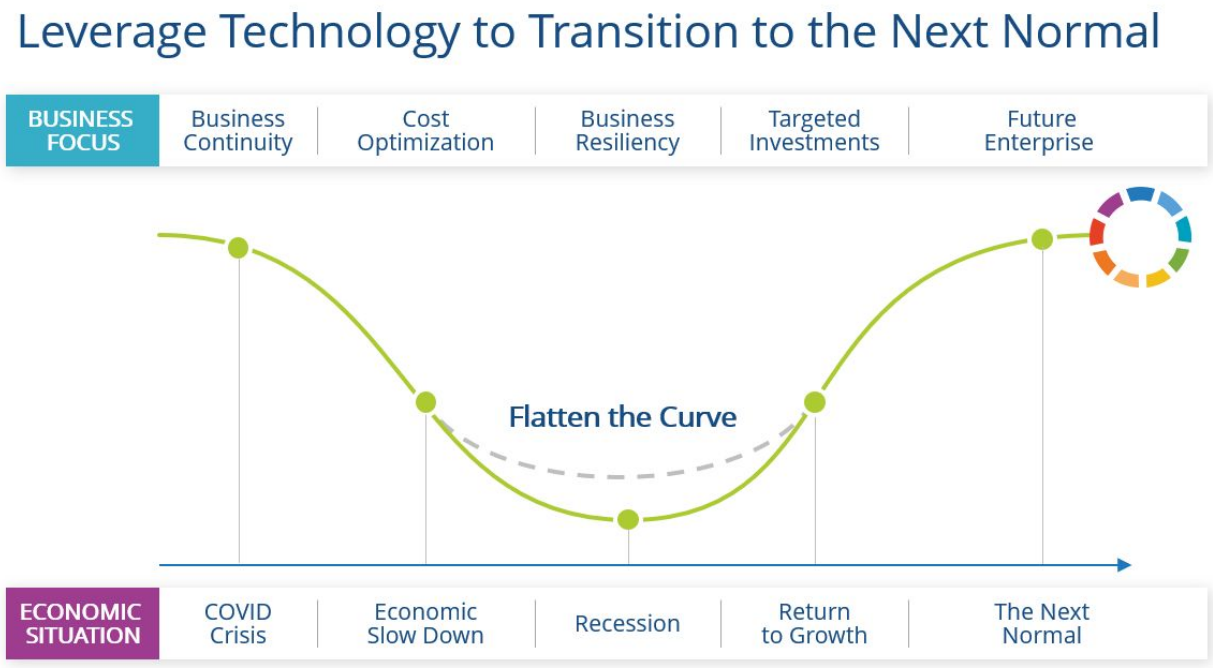


Fig 2. Top 4 Affected Countries



4. Methodology

Most of the research papers and articles focus on the COVID-19 infection in the entire India. But considering the size and diversity of our country , it would be a good idea to look at the spread of the virus in each state separately along with the entire nation. We use a modified version of the mathematical model **SEIR** and the **Prophet** model for predicting the infection. It was reported that the growth of COVID-19 infection would be controlled with the help of National Lockdown, but due to carelessness, negligence and some uncontrolled events there has been an exponential rise in the COVID cases throughout the nation. In this project we have used exponential and polynomial regression modelling to make predictions of cases till November 2020.

3.1 Dataset Description:

For the analysis and prediction of COVID-19 infection, we mainly considered 3 datasets : First two datasets contain cases in India : state_wise distribution and district_wise distribution. These two datasets are taken from COVID19-India API. They contain the following information : number of confirmed cases, recovered cases and deceased and also number of tests per day. The first dataset contains these information of each state while the second contains information of each district. The third dataset is taken from covid.ourworldindata.org It contains world data for covid cases. It contains all the important information such as new cases, total cases, total deaths, etc.

3.2 Mathematical Modelling

- **SEIR Model :** We generated a modified SEIRS epidemic model for COVID-19 cases in India. In a closed population without births or deaths, the classical SEIR model is:

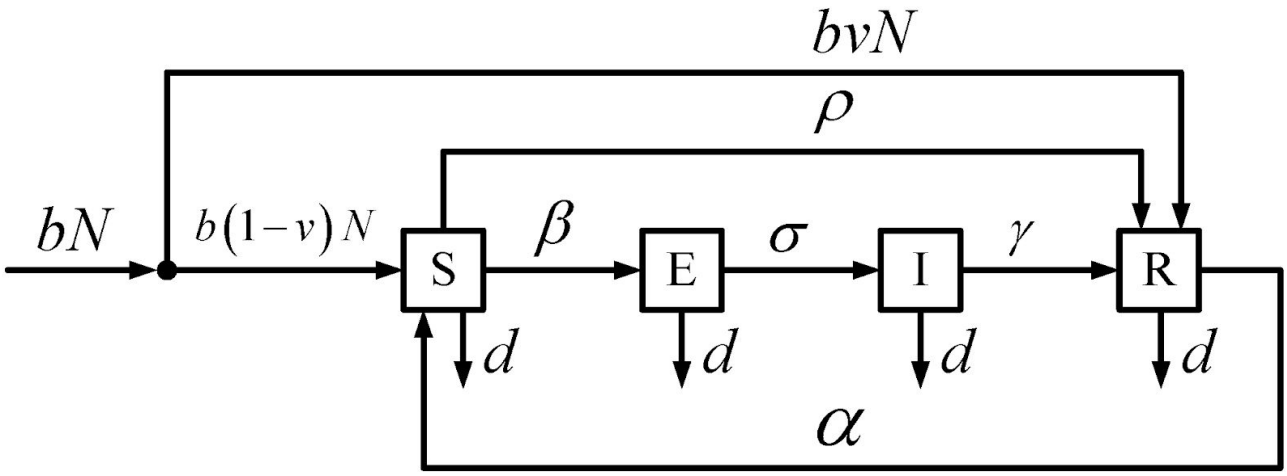
$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI + \xi R$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\beta SI}{N} - \sigma E$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \xi R$$

where S, E, I, R is the proportion of susceptible, exposed, infectious and recovered populations.



In the case of COVID-19, the viral carriers ("exposed population) do not exhibit symptoms, yet are infectious. So, the SEIR model is modified as below:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\frac{\beta S(I+E)}{N} + \xi R$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\beta S(I+E)}{N} - \sigma E$$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = \sigma E - \gamma I$$

$$\frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \xi R$$

Where S, E, I, R is the proportion of susceptible, exposed, infectious, and recovered population.

We use our model to predict new cases, recovered cases and deceased cases of COVID-19 in India.

- **Prophet Model:** It is a popular model developed by Google which is used for forecasting time series data. It is Simply based on an additive model. It can fit Linear or non-linear trends in data with regular intervals. This is similar to an additive model, with time as a regressor. It has three main model components: pattern, seasonality, and holidays, it makes use of a decomposable time series model. This is analogous to a model of additives, with time as a regressor. The Prophet integrates as components several linear and nonlinear functions of time. In its general form:

$$y(t) = g(t) + s(t) + h(t) + e(t)$$

where:

g(t) : trend models non-periodic changes

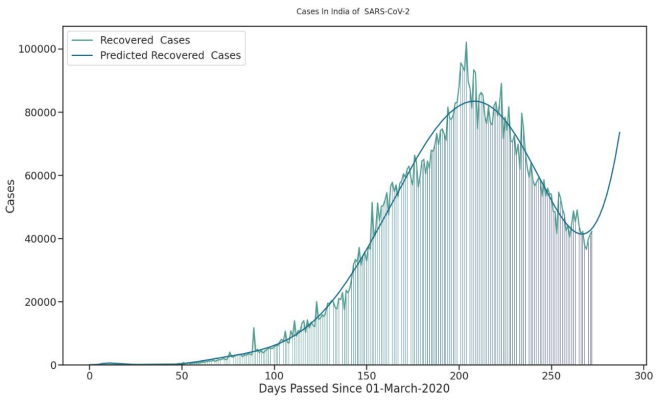
s(t) : seasonality models periodic changes

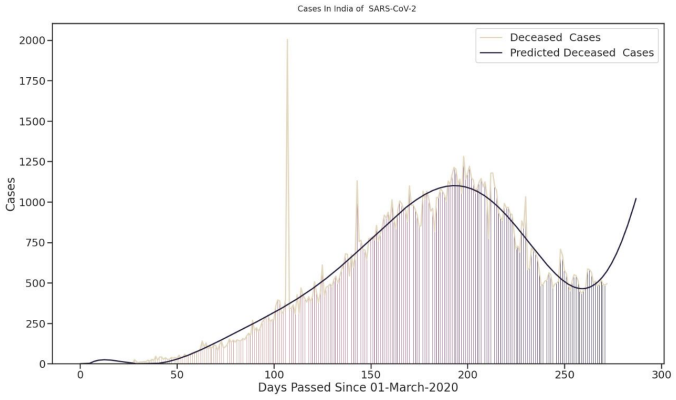
h(t) : ties in effects of holidays (on potentially irregular schedules ⩾ 1 day(s))

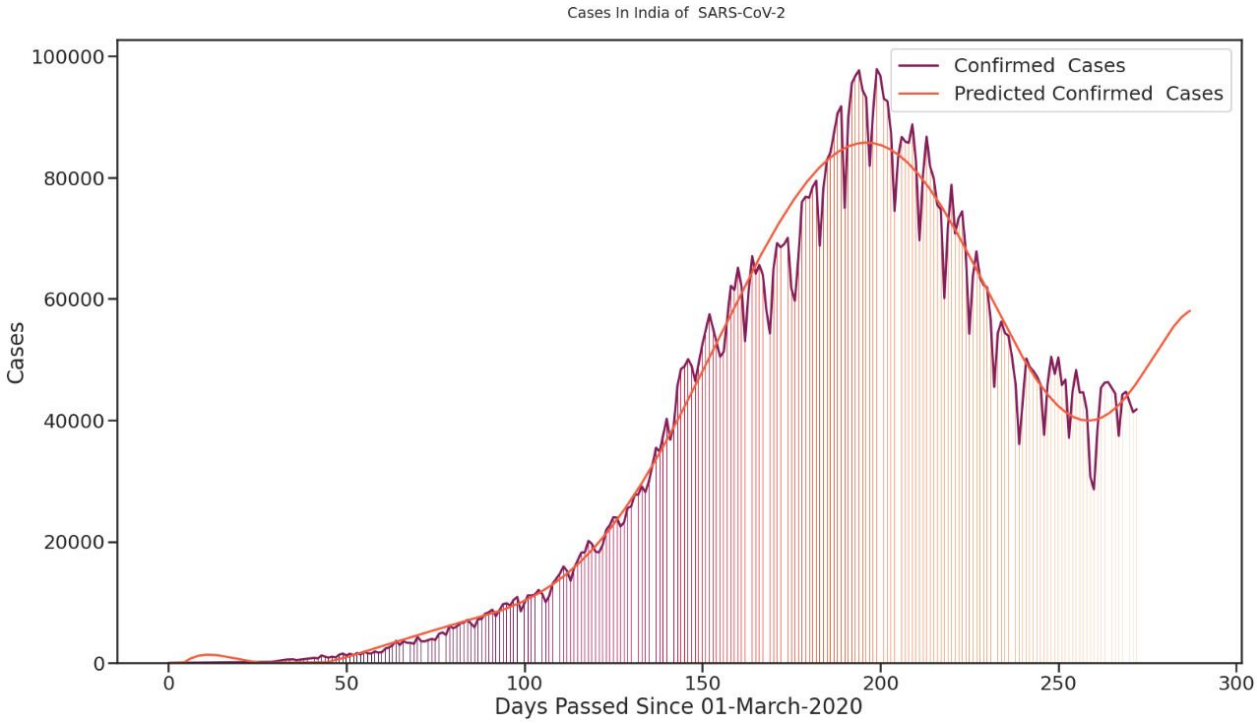
e(t) : it covers idiosyncratic changes not accommodated by the model

3.3 Regression Predictions:

We used polynomial regressions for predictions on confirmed cases, recovered cases and deceased cases. It is clear from the graph that cases were very less in the initial days but after 90 days (after lockdown was revoked), new cases have increased exponentially.

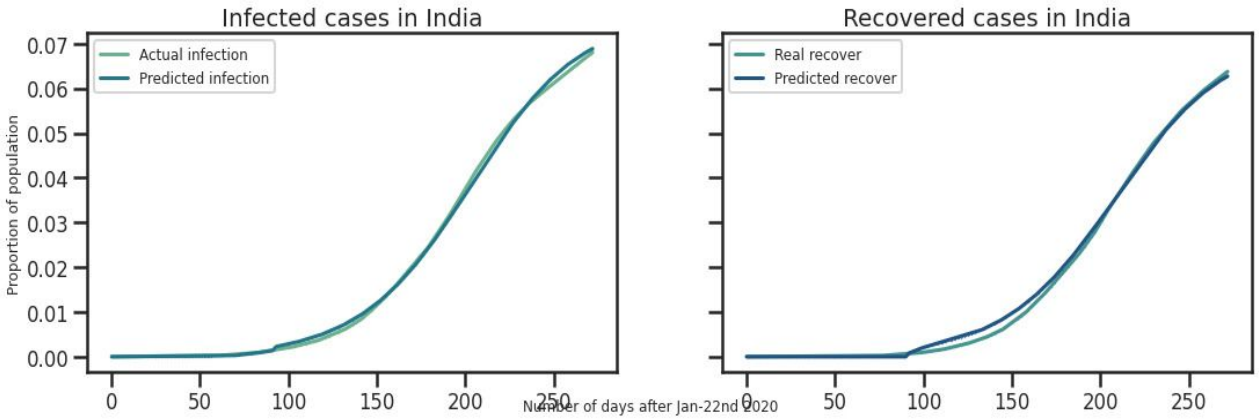




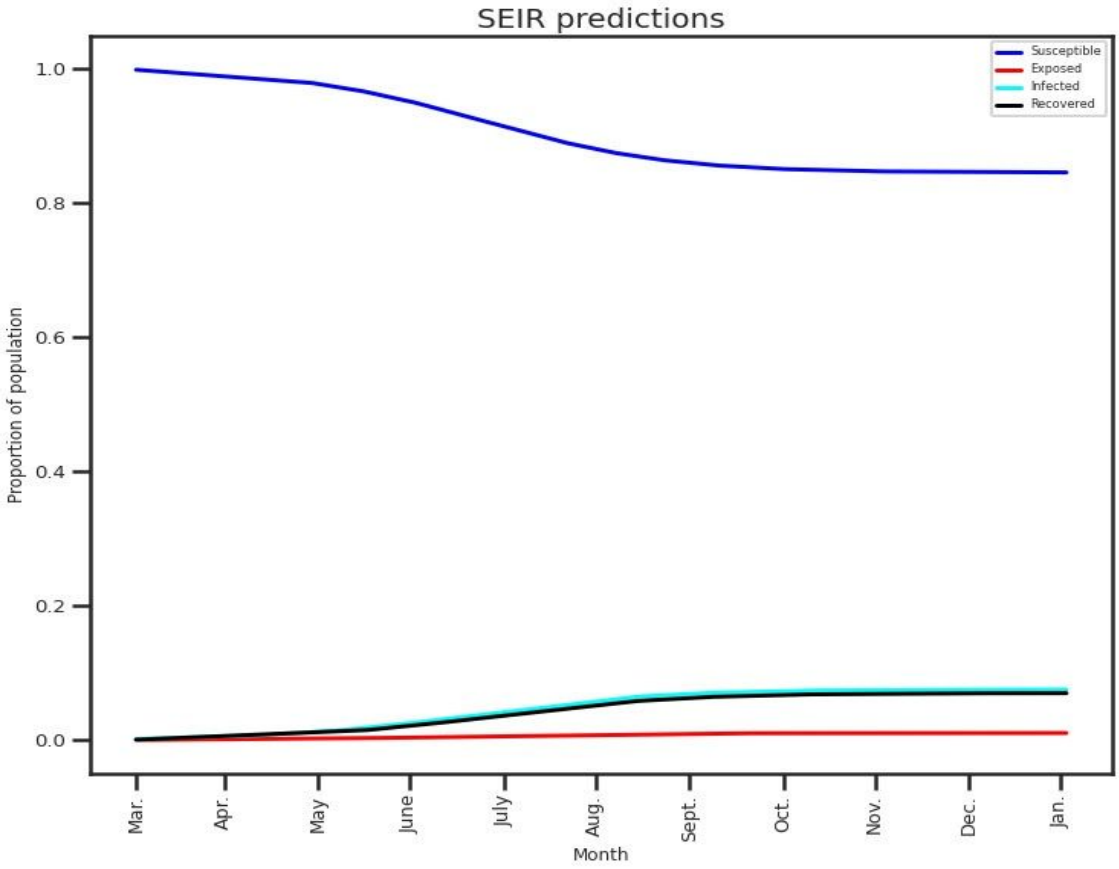


5. Result

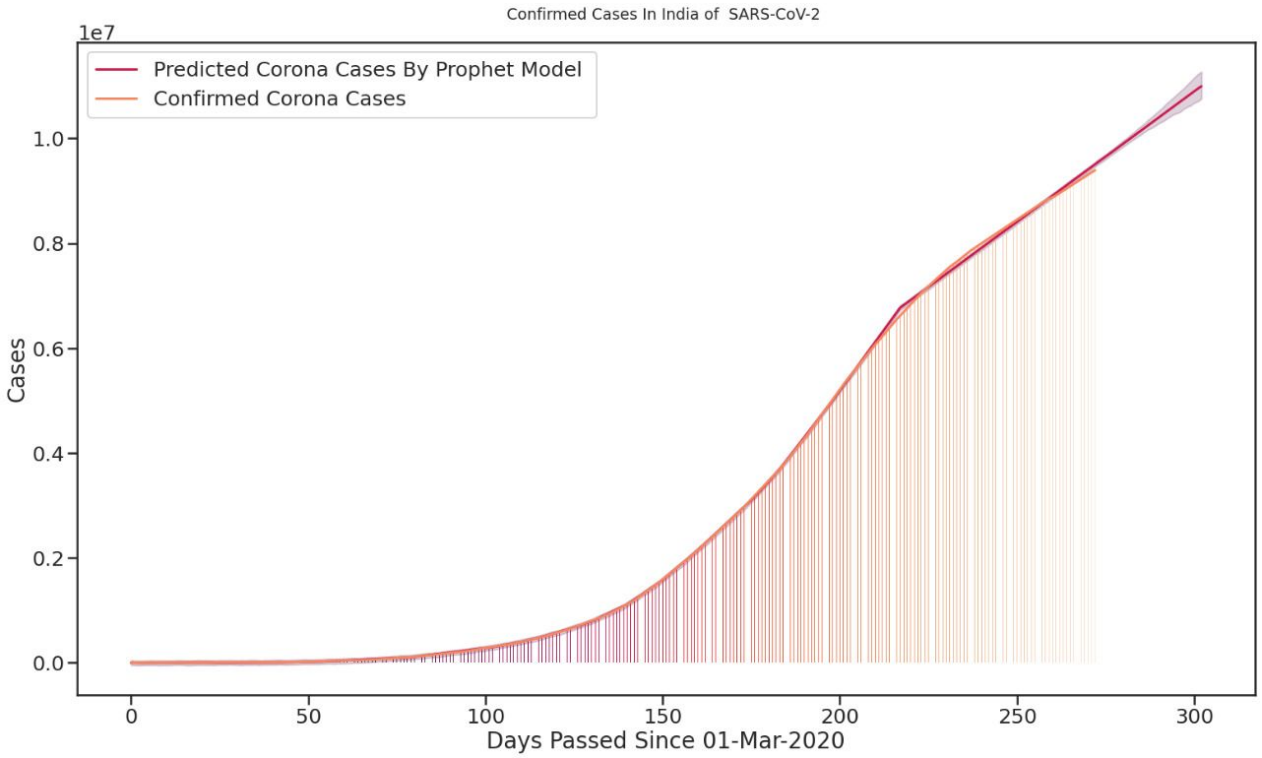
- **<u>The results for the SEIR model are as follows:</u>**



Our model performs well in predicting the COVID-19 infection in India. The above graph shows the comparison between actual infection vs predicted infection cases and actual recovered vs predicted recovered infection cases.



- **<u>The results of the Prophet model are as follows:</u>**



Our model performs well in predicting the COVID-19 infection in India. The above graph shows the comparison between actual confirmed cases vs the predicted cases. Its clear from the graph after 90 days the growth in no. of cases is exponential .

4.1 Conclusion

We present a comprehensive analysis of the COVID-19 infection in India. Although the cases were less in the initial days but they have risen abruptly in the last 2-3 months. Our work demonstrates growth patterns in infected cases in India, estimates of the number of infected cases for the next few days, the effect of social distancing on Indian people, the impact of mass events on the number of infected cases in India, network analysis and the analysis of strategies for uplifting lockdown in India. The cases are rising very fast, although now there has been some decay in the growth rate of infections but still there are aound 40-45k new cases seen in India. We have to take proper measures and maintain social distancing and proper hygiene to tackle this pandemic.